Part 1: Introduction to Data

Sam Tyner

TBD

Textbook

These slides are based on the book *OpenIntro Statistics* by David Diez, Christopher Barr, and Mine Çetinkaya-Rundel

The book can be downloaded from https://www.openintro.org/stat/textbook.php

Part 1 Corresponds to Chapter 1 of the text. Sections 1.1-1.7 correspond to sections 1.1-1.7 of the text.

Outline

- Introductory example - why statistics? (1.1)
- Data basics (1.2)
- Overview of data collection principles (1.3)
- Obervational studies and sampling strategies (1.4)
- Experiments (1.5)
- Looking at Numerical Data (1.6)
- Looking at Categorical data (1.7)

Section 1.1: Introductory example - Why Statistics?

Why Statistics?

- Good scientists use rigorous methods and make careful observations
- Observations aka **data**: the backbone of a statistical investigation

Why Statistics?

- Good scientists use rigorous methods and make careful observations
- Observations aka **data**: the backbone of a statistical investigation
- **Statistics**: "the study of how best to collect, analyze, and draw conclusions from data" (from the textbook)

Why Statistics?

- Good scientists use rigorous methods and make careful observations
- Observations aka **data**: the backbone of a statistical investigation
- **Statistics**: "the study of how best to collect, analyze, and draw conclusions from data" (from the textbook)
- **Statistics**: "a branch of mathematics dealing with the collection, classification, analysis, interpretation of numerical facts, for drawing inferences on the basis of their quantifiable likelihood (probability) of data." (from Wikipedia)

Why Statistics?

- Good scientists use rigorous methods and make careful observations
- Observations aka **data**: the backbone of a statistical investigation
- **Statistics**: "the study of how best to collect, analyze, and draw conclusions from data" (from the textbook)
- **Statistics**: "a branch of mathematics dealing with the collection, classification, analysis, interpretation of numerical facts, for drawing inferences on the basis of their quantifiable likelihood (probability) of data." (from Wikipedia)
- **Statistics**: "the study of variation" (my definition)

Statistics in context

General investigative process:

1. Identify a question or problem.
2. Collect relevant data on the topic.
3. Analyze the data.
4. Form a conclusion.

Where do you think “statistics” fit in?

Statistics in context

General investigative process:

1. Identify a question or problem.
2. Collect relevant data on the topic.
3. Analyze the data.
4. Form a conclusion.

Statistics focuses on making stages 2-4 objective, rigorous, and efficient

The three pieces of statistics

1. How *best* can we collect data?
2. How *should* it be analyzed?
3. And what can we *infer* from the analysis?

Classical application: medical trials

1. Identify a question or problem.
2. Collect relevant data on the topic.
3. Analyze the data.
4. Form a conclusion.

1. Identify a question or problem.

Does the use of stents reduce the risk of stroke?

Will patients with stents inserted have better outcomes (fewer and/or more minor strokes) than patients without stents inserted?

Stents - “a metal or plastic tube inserted into the lumen of an anatomic vessel or duct to keep the passageway open” (Wikipedia)

2. Collect relevant data on the topic.

451 at-risk patients volunteered to be studied. Split into 2 groups:

- **Treatment** group (224 patients): Received a stent and medical management (medications, management of risk factors, and help in lifestyle modification)
- **Control** group (227 patients): Received same medical management as treatment group, but did not receive stents.

What is the purpose of the control group?

2. Collect relevant data on the topic.

451 at-risk patients volunteered to be studied. Split into 2 groups:

- **Treatment** group (224 patients): Received a stent and medical management (medications, management of risk factors, and help in lifestyle modification)
- **Control** group (227 patients): Received same medical management as treatment group, but did not receive stents.

What is the purpose of the control group?

The control group provides a reference point against which we can measure the medical impact of stents in the treatment group.

2. Collect relevant data on the topic. (cont.)

- Researchers looked at 2 time points: 30 days after enrollment and 365 days after enrollment
- Results of 5 patients:

Patient	group	after 30 days	after 365 days
1	treatment	no event	no event
2	treatment	stroke	stroke
3	treatment	no event	no event
. . .	. . .	. . .	. . .
450	control	no event	no event
451	control	no event	no event

2. Collect relevant data on the topic. (cont.)

All outcomes in all patients in all groups:

group	timepoint	outcome	count
treatment	0-30 days	stroke	33
treatment	0-30 days	no event	191
treatment	0-365 days	stroke	45
treatment	0-365 days	no event	179
control	0-30 days	stroke	13
control	0-30 days	no event	214
control	0-365 days	stroke	28
control	0-365 days	no event	199

3. Analyze the data.

Summary statistic - one numerical value that summarizes a large amount of data

What do you think would be a good summary statistic for this data?

3. Analyze the data.

Summary statistic - one numerical value that summarizes a large amount of data

What do you think would be a good summary statistic for this data?
Proportion of patients having a stroke

- Proportion of patients who had a stroke in the treatment group: $45/224 = 0.20 = 20\%$.
- Proportion of patients who had a stroke in the control group: $28/227 = 0.12 = 12\%$.
- Overall proportion of patients who had a stroke: $73/451 = 0.16 = 16\%$

4. Form a conclusion.

Surprising result: treatment group has higher rate of stroke (8% more)

- Not what doctors expect
- Is this difference meaningful?

Not yet equipped with statistical tools to answer the 2nd question: is the difference between stent and non-stent groups so large that we should reject the notion that it was due to random variation?

We'll get there eventually!

Your Turn 1.1.1

Researchers studying the effect of temperature on glass looked at the refractive index (RI) measured on 89 different types of glass before and after being exposed to extreme temperatures. The RI value on all glass types was recorded, then the glass was exposed to very cold (0^oF) or very hot (400^oF) conditions for 20 minutes. After the glass returned to room temperature, the researchers measured the RI again to see if it was the same or if it had changed. Results are summarized below:

Group	change	no change	Total
Heat	10	33	43
Cold	2	44	46
Total	12	77	89

Your Turn 1.1.1 (cont.)

Group	change	no change	Total
Heat	10	33	43
Cold	2	44	46
Total	12	77	89

1. What percent of observations had a change in RI after being exposed to extreme heat? What percent of observations had a change in RI after being exposed to extreme cold?
2. At first glance, does exposure to extreme temperatures appear to have an effect on RI of glass? Explain.
3. Do the data provide convincing evidence that there is a real change in RI for either group?
4. Do you think the effect may just be due to random variability (as opposed to temperature)?

Your Turn 1.1.1 (soln.)

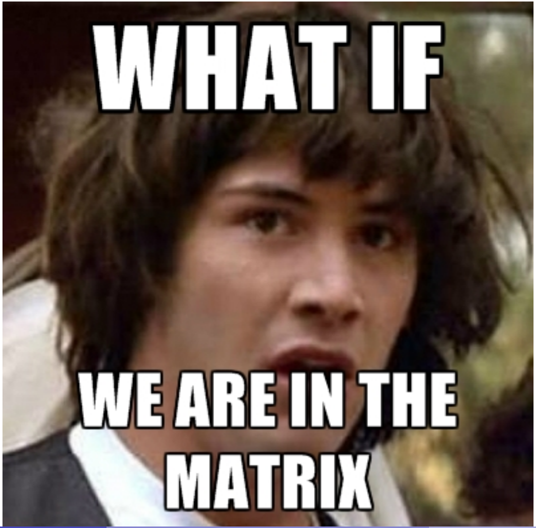
Group	change	no change	Total
Heat	10	33	43
Cold	2	44	46
Total	12	77	89

1. What percent of observations had a change in RI after being exposed to extreme heat? $\frac{10}{43} = 0.23 = 23\%$ What percent of observations had a change in RI after being exposed to extreme cold? $\frac{2}{46} = 0.04 = 4\%$
2. At first glance, does exposure to extreme temperatures appear to have an effect on RI of glass? Explain. Heat, yes. Cold, no.
3. Do the data provide convincing evidence that there is a real change in RI for either group? (discussion)
4. Do you think the effect may just be due to random variability (as opposed to temperature)? (discussion)

Section 1.2: Data basics

Observations, variables, and matrices

Data are stored in **tables** aka **matrices**



Observations, variables, and matrices

Data are stored in **tables** aka **matrices**:

- each row is an **observation**
- each column is a **variable**
- data are stored in a **matrix**
- Example: 6 randomly selected obervations of 6 variables on glass element analysis. (units of last 3 columns are ppm)

pane	Piece	Rep	Li7	Na23	Mg25
P4	2	6	2.23	105190	24580
P3	16	5	3.79	108500	24910
P4	14	15	1.42	100860	25080
P1	7	3	1.83	102160	23060
P2	17	3	1.35	99170	23810
P2	21	4	1.67	100300	23370

Types of variables

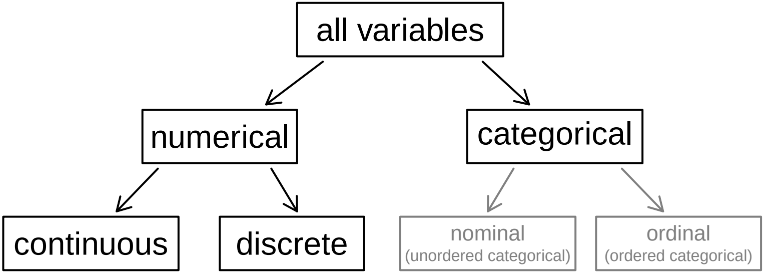


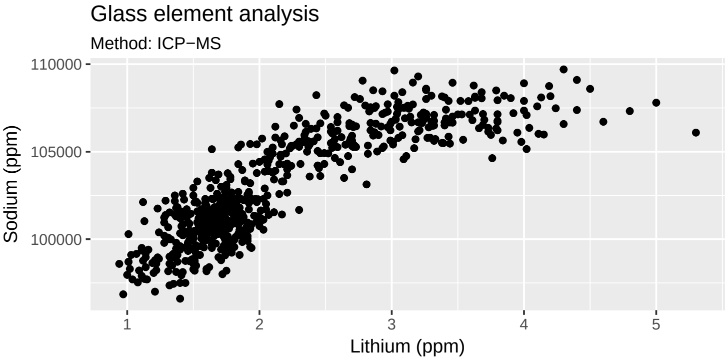
Figure 2: Types of variables

Link to Image Source (from textbook)

Relationships between variables

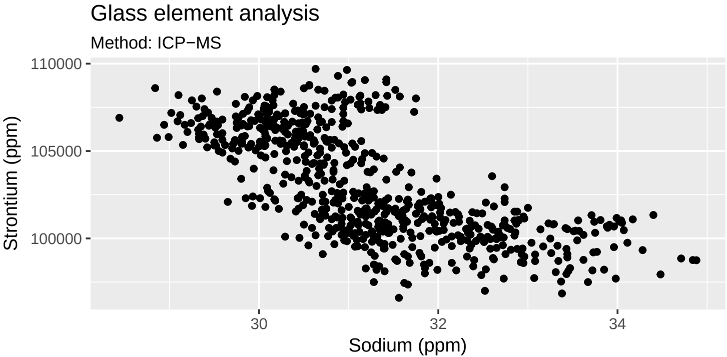
Most (if not all) analyses are performed in order to determine if there is a relationship between variables

Is there a relationship between these two variables in the glass data?



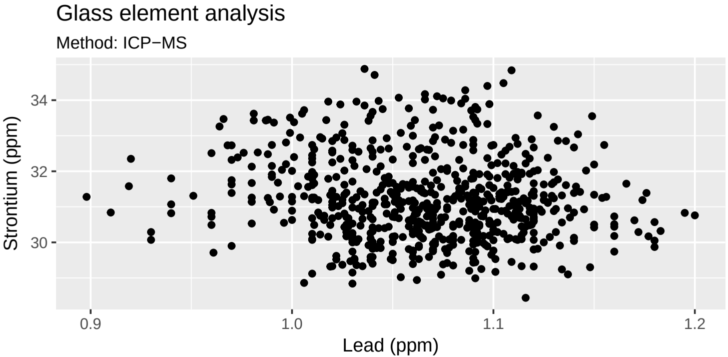
Relationships between variables

Is there a relationship between these two variables in the glass data?



Relationships between variables

Is there a relationship between these two variables in the glass data?



Relationships between variables

- **positive** association (correlation) between 2 variables
- **negative** association (correlation) between 2 variables
- **no** association between two variables

Relationships between variables

- **Positive** association: Values of one variable tend to **increase** when values of the other variable increase
- **Negative** association: Values of one variable tend to **decrease** when values of the other variable increase
- **No** association: **Independence** occurs when there is no relationship between two variables

Your Turn 1.2.1

Researchers collected data to examine the relationship between air pollutants and preterm births in Southern California. During the study air pollution levels were measured by air quality monitoring stations. Levels of CO were recorded in ppm, NO_2 and O_3 in pphm, and coarse particulate matter (PM_{10}) in $\mu g/m^3$. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth. The analysis suggested that increased ambient PM_{10} and, to a lesser degree, CO concentrations may be associated with the occurrence of preterm births. In this study, identify:

1. The observational units
2. The variables and their types
3. The main research question

Your Turn 1 1.2.1 (soln.)

Researchers collected data to examine the relationship between air pollutants and preterm births in Southern California. During the study air pollution levels were measured by air quality monitoring stations. Levels of CO were recorded in ppm, NO_2 and O_3 in pphm, and coarse particulate matter (PM_{10}) in $\mu g/m^3$. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth. The analysis suggested that increased ambient PM_{10} and, to a lesser degree, CO concentrations may be associated with the occurrence of preterm births. In this study, identify:

1. The observational units - 143,196 births in Southern California from 1989-1993
2. The variables and their types - Length of gestation, CO, NO_2, O_3, and PM_{10}. All are numerical & continuous
3. The main research question - Is there an association between pregnant mothers' exposure to air pollution and preterm births?

Your Turn 1.2.2

A survey was conducted to study the smoking habits of UK residents. Below is a data matrix displaying a portion of the data collected in this survey. Note that NA refers to a missing component of the data.

	gender	age	maritalStatus	grossIncome	smoke	amtWeekends	amtWeekdays
1	Male	38	Divorced	2,600 to 5,200	No	NA	NA
2	Female	42	Single	Under 2,600	Yes	12	12
3	Male	40	Married	28,600 to 36,400	No	NA	NA
1691	Male	31	Married	10,400 to 15,600	No	NA	NA

1. What does each row of the data matrix represent?
2. How many participants were included in the survey?
3. Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

Your Turn 1.2.2 (soln.)

A survey was conducted to study the smoking habits of UK residents. Below is a data matrix displaying a portion of the data collected in this survey. Note that NA refers to a missing component of the data.

	gender	age	maritalStatus	grossIncome	smoke	amtWeekends	amtWeekdays
1	Male	38	Divorced	2,600 to 5,200	No	NA	NA
2	Female	42	Single	Under 2,600	Yes	12	12
3	Male	40	Married	28,600 to 36,400	No	NA	NA
1691	Male	31	Married	10,400 to 15,600	No	NA	NA

1. What does each row of the data matrix represent? a person living in the UK
2. How many participants were included in the survey? 1691
3. Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal. Numerical: age, amtWeekends, amtWeekdays (discrete). Categorical: gender, marital, grossIncome (ordinal), smoke

Section 1.3: Overview of data collection principles

Populations and samples

Consider 3 research questions

1. What is the average lead content in float glass manufactured in the US?
2. Since 2010, what is the average length of the training process to become a forensic scientist in the US?
3. Does a new powder lift up higher quality latent prints than the standard powder?

For each question: What is the **population** of interest? What would a **sample** be?

Populations and samples

1. **Population**: All float glass manufactured in the US. **Sample**: The windows in Durham Center.
2. **Population**: All forensic scientists who started training in 2010 or later. **Sample**: The new and recent hires at Iowa's DCI Lab
3. **Population**: All latent prints. **Sample**: Some latent prints picked up in this conference room.

Summary

Population: all items you're interested in for your research questions

Sample: the subset of items in the population that you collect data on

Anecdotal evidence

Anecdote = story

One or two or a handful of observations cannot be proof of a pattern

Haphazard, cherry-picked examples ≠ Science!

Sampling from a population

Sampling should be **random**

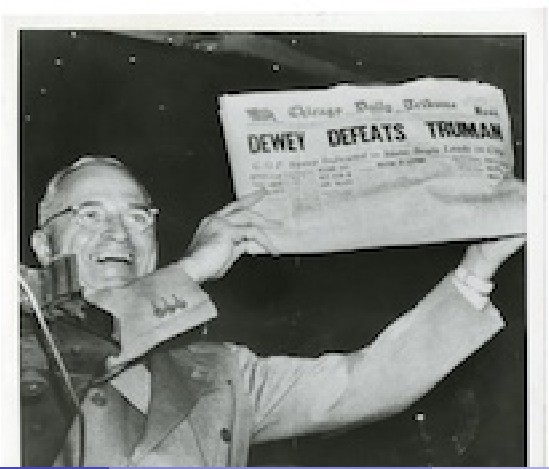
- think about a raffle or roll of a die
- every person has one ticket, every number has one side of the die

Sampling should be **representative**

- all elements of the population should have a chance to be in the sample

Where sampling goes wrong

- **bias**: the method of collection favors one type of response over others. Example:



Where sampling goes wrong

- **non-response**: is a survey where only 20% of people respond *actually* representative?

Example: survey on drug use in teenagers

- **convenience sample**: only sample people/things that are most accessible. Sample is not represenative.

Example: How much do people spend on groceries every week? Standing outside of Whole Foods will get very different answers than standing outside of Wal-Mart

Simple random sampling

Simple random sampling (SRS) is the best way to sample in most cases

- every element of the population could be selected
- every element of the poulation has the same chance of being selected

Explanatory and response variables

response variable: the main thing you want to know about your data

What is the response variable in the 3 scenarios from earlier?

1. What is the average lead content in float glass manufactured in the US?
2. Since 2010, what is the average length of the training process to become a forensic scientist in the US?
3. Does a new powder lift up higher quality latent prints than the standard powder?

Explanatory and response variables

response variable: the main thing you want to know about your data

What is the response variable in the 3 scenarios from earlier?

1. What is the average lead content in float glass manufactured in the US? Lead content measured in glass
2. Since 2010, what is the average length of the training process to become a forensic scientist in the US? time a forensic scientist spent in training (in months)
3. Does a new powder lift up higher quality latent prints than the standard powder? print quality/clarity

Explanatory and response variables

explanatory variable: may have an effect on the response variable, and you want to study the relationship between the response and expanatory variables

What is the explanatory variable in the 3 scenarios from earlier (if applicable)?

1. What is the average lead content in float glass manufactured in the US?
2. Since 2010, what is the average length of the training process to become a forensic scientist in the US?
3. Does a new powder lift up higher quality latent prints than the standard powder?

Explanatory and response variables

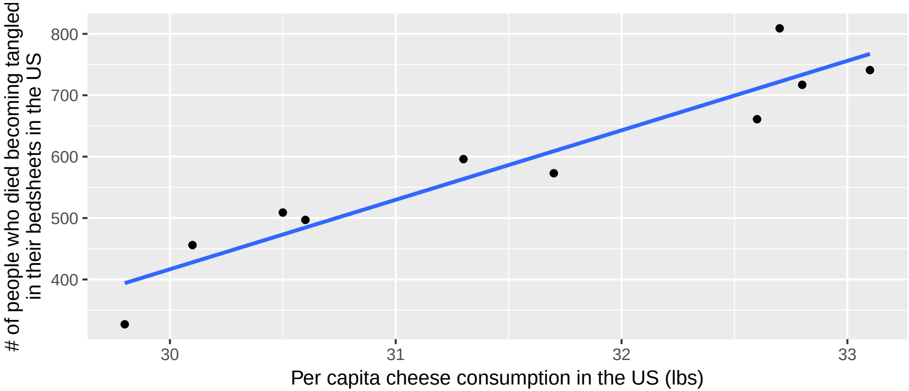
explanatory variable: may have an effect on the response variable, and you want to study the relationship between the response and expanatory variables

What is the explanatory variable in the 3 scenarios from earlier (if applicable)?

1. What is the average lead content in float glass manufactured in the US? (discussion)
2. Since 2010, what is the average length of the training process to become a forensic scientist in the US? (discussion)
3. Does a new powder lift up higher quality latent prints than the standard powder? powder type (new or old)

Correlation $\neq$ Causation

An association between the explanatory and response variable does not always mean that the explanatory variable causes the change in the response variable.



Causation: only from experiments

- To determine a causal relationship between and explanatory and response variable, an *experiment* must be done
- Often unethical or impossible to do an experiment (e.g. smoking and lung cancer), so an **observational study** is a substitute

Your Turn 1.3.1

Recall the study where researchers collected data to examine the relationship between air pollutants and preterm births in Southern California.

1. Identify the population of interest and the sample in this study.
2. Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

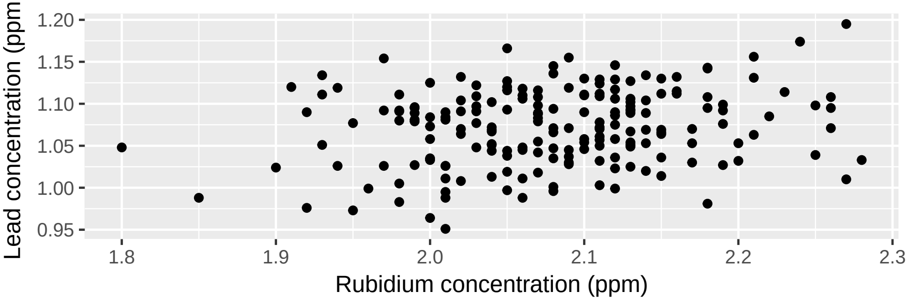
Your Turn 1.3.1 (soln.)

Recall the study where researchers collected data to examine the relationship between air pollutants and preterm births in Southern California.

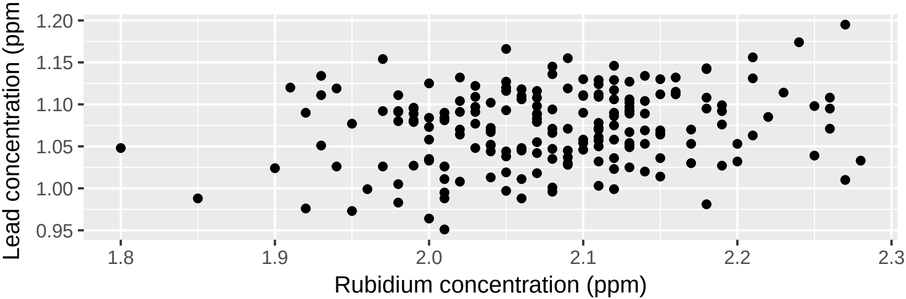
1. Identify the population of interest and the sample in this study.
   Population: all births. Sample: 143,196 births between 1989-1993 in Southern California
2. Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships. (discussion)

Your Turn 1.3.2

Researchers are interested in predicting the amount of lead in a pane of glass by measuring the content of another element that is much cheaper to measure, rubidium. A random sample of 193 panes of glass from homes built before 1950 in Ames, Iowa was taken. The scatterplot below displays the relationship between these two elements.



Your Turn 1.3.2 (cont.)



1. What is the explanatory variable and what is the response variable?
2. Describe the relationship between the two variables.
3. Is this an experiment or an observational study?
4. Can we conclude that more rubidium in the glass means more lead in the glass?

Your Turn 1.3.2 (soln.)

1. What is the explanatory variable and what is the response variable? Explanatory variable: rubidium concentration; Respons variable: lead concentration
2. Describe the relationship between the two variables. There is a slight positive association between the two, meaning that lead content may increase with rubidium content.
3. Is this an experiment or an observational study? Observational (no control)
4. Can we conclude that more rubidium in the glass means more lead in the glass? No, because we can not infer a causal relationship without an experiment.

Section 1.4: Obervational studies and sampling strategies

Observational studies

- researchers *observe* the data
- no intereference
- there may be **counfounding** variables that are correlated with both the explanatory and response variables

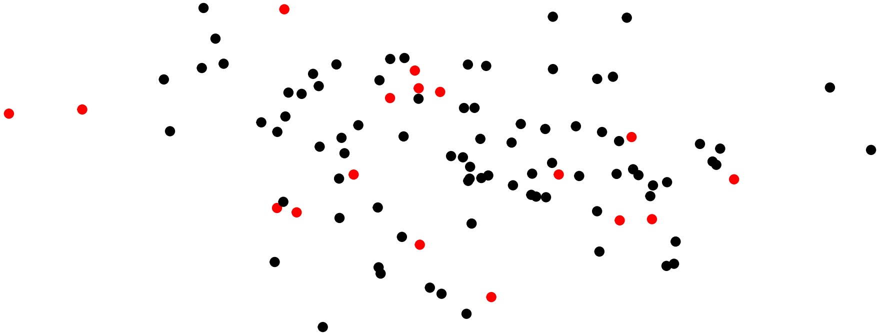
Sampling strategies

- **Simple random sampling**
- **Stratified sampling**
- **Cluster sampling**
- **Multistage sampling**
- **Census**

Simple random sampling

- all elements have the same chance of being chosen
- the selection of one element does not affect the chances of a second element being selected

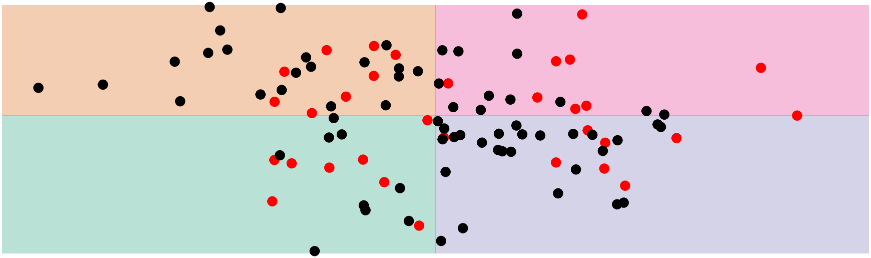
Simple random sample



Stratified sampling

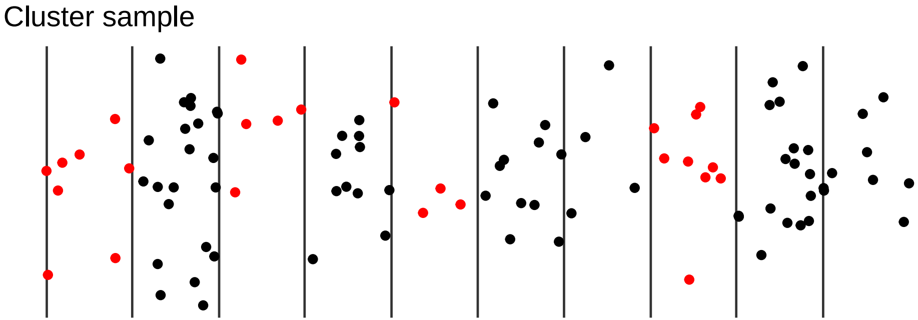
- Population is made up of different groups or **strata**
- “divide and conquer”
- elements within a strata are more similar than elements from different strata
- draw simple random samples from each strata

Stratified sample



Cluster sampling

- Divide population into representative **clusters**
- Clusters are similar
- Randomly sample clusters, and then take a census of all elements in the cluster

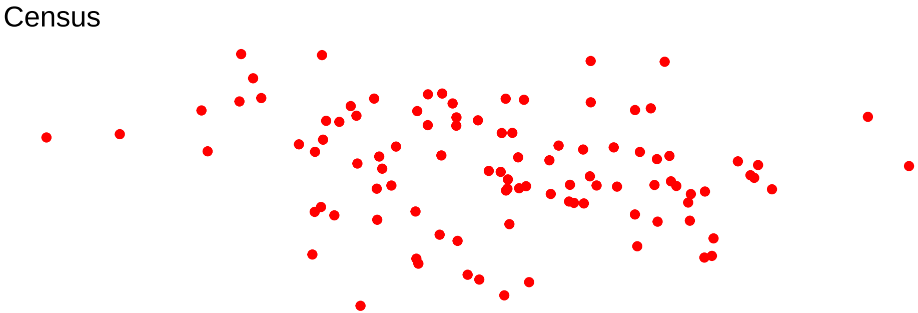


Multistage sampling

- Like cluster sampling, but take a random sample from the clusters instead of a census
- Done when getting an observation is very expensive

Census

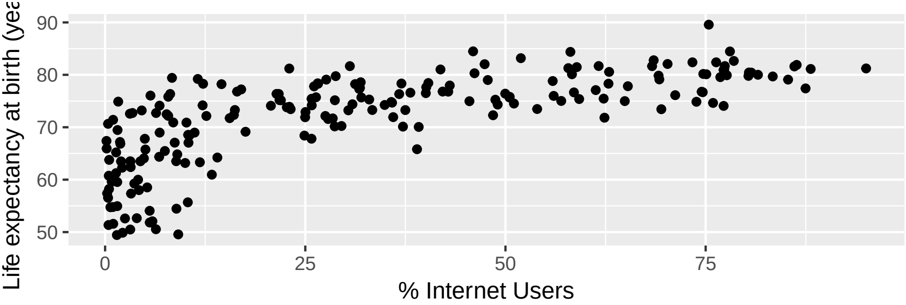
- Observe all elements in a population
- Rarely done because it's so expensive



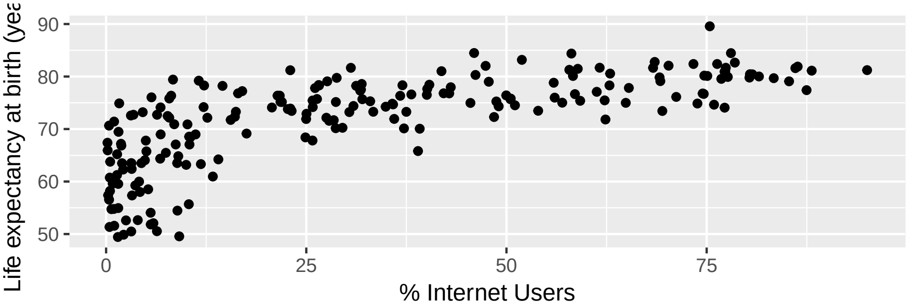
Census

Your Turn 1.4.1

The following scatterplot was created as part of a study evaluating the relationship between estimated life expectancy at birth (as of 2014) and percentage of internet users (as of 2009) in 208 countries for which such data were available.

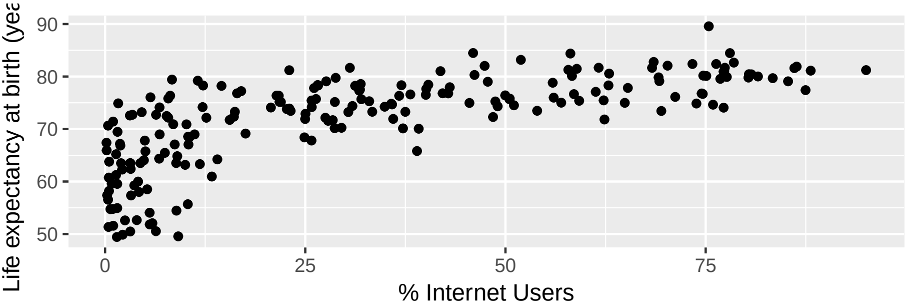


Your Turn 1.4.1 (cont.)



1. Describe the relationship between life expectancy and percentage of internet users.
2. What type of study is this?
3. State a possible confounding variable that might explain this relationship and describe its potential effect.

Your Turn 1.4.1 (soln.)



1. Describe the relationship between life expectancy and percentage of internet users. Positive, non-linear relationship. Levels off at around age 75
2. What type of study is this? Observational
3. State a possible confounding variable that might explain this

Your Turn 1.4.2

Identify the flaw(s) in reasoning in the following scenarios. What should have been done differently to arrive at the desired conclusions?

1. Students at an elementary school are given a questionnaire that they are asked to return after their parents have completed it. One of the questions asked is, "Do you find that your work schedule makes it difficult for you to spend time with your kids after school?" Of the parents who replied, 85% said "no". Based on these results, the school officials conclude that a great majority of the parents have no difficulty spending time with their kids after school.

2. A survey is conducted on a simple random sample of 1,000 crime victims (robbery, assault, etc.) asking whether or not they saught mental health counseling after their crime. A follow-up survey asking asking the same question 3 years later is condcuted, however, only 567 of these original participants are reached at the same address. The researcher reports that these 567 people are representative of all crime victims.

Your Turn 1.4.2 (soln.)

Identify the flaw(s) in reasoning in the following scenarios. What should have been done differently to arrive at the desired conclusions?

1. Students at an elementary school are given a questionnaire that they are asked to return after their parents have completed it. One of the questions asked is, "Do you find that your work schedule makes it difficult for you to spend time with your kids after school?" Of the parents who replied, 85% said "no". Based on these results, the school o cials conclude that a great majority of the parents have no difficulty spending time with their kids after school. Parents with time to fill out a survey probably have more time to spend with their children after work. (discussion)

2. A survey is conducted on a simple random sample of 1,000 crime victims (robbery, assault, etc.) asking whether or not they saught mental health counseling after their crime. A follow-up survey asking asking the same question 3 years later is condcuted, however, only 567 of these original participants are reached at the same address. The researcher reports that these 567 people are representative of all crime victims. People who moved away may have felt unsafe, perhaps they would not have felt unsafe if they saught counseling. (discussion)

Your Turn 1.4.3

Below is an excerpt from an article published in the NY Times:

“Risks: Smokers Found More Prone to Dementia" (link):

“Researchers analyzed the data of 23,123 health plan members who participated in a voluntary exam and health behavior survey from 1978 to 1985, when they were 50 to 60 years old. 23 years later, about 1/4 of the group, or 5,367, had dementia, including 1,136 with Alzheimer’s disease and 416 with vascular dementia. After adjusting for other factors, the researchers concluded that pack-a-day smokers were 37% more likely than nonsmokers to develop dementia, and the risks went up sharply with increased smoking; 44% for 1-2 packs a day; and twice the risk for more than 2 packs."

Based on this study, can we conclude that smoking causes dementia later in life? Explain.

Your Turn 1.4.3 (soln.)

Below is an excerpt from an article published in the NY Times:

“Risks: Smokers Found More Prone to Dementia" (link):

“Researchers analyzed the data of 23,123 health plan members who participated in a voluntary exam and health behavior survey from 1978 to 1985, when they were 50 to 60 years old. 23 years later, about 1/4 of the group, or 5,367, had dementia, including 1,136 with Alzheimer’s disease and 416 with vascular dementia. After adjusting for other factors, the researchers concluded that pack-a-day smokers were 37% more likely than nonsmokers to develop dementia, and the risks went up sharply with increased smoking; 44% for 1-2 packs a day; and twice the risk for more than 2 packs."

Based on this study, can we conclude that smoking causes dementia later in life? Explain.

Association $\neq$ causation. Observational study, not experiment. However, experiment would be unethical in this case.

Your Turn 1.4.4

Below is another excerpt from another article published in the NY Times:

“The School Bully is Sleepy" (link):

“The University of Michigan study collected data from parents on each child’s sleep habits and asked both parents and teachers to assess behavioral concerns. Among the 341 children studied, about a third were identified by parents or teachers as having problems with disruptive behavior or bullying. The researchers found that children who had behavioral issues were twice as likely to have shown symptoms of sleep-disordered breathing, like snoring or daytime sleepiness."

A friend of yours who read the article says, “The study shows that sleep disorders lead to bullying in school children." Is this statement justified? If not, how best can you describe the conclusion that can be drawn from this study?

Your Turn 1.4.4 (soln.)

Below is another excerpt from another article published in the NY Times:

“The School Bully is Sleepy" (link):

“The University of Michigan study collected data from parents on each child’s sleep habits and asked both parents and teachers to assess behavioral concerns. Among the 341 children studied, about a third were identified by parents or teachers as having problems with disruptive behavior or bullying. The researchers found that children who had behavioral issues were twice as likely to have shown symptoms of sleep-disordered breathing, like snoring or daytime sleepiness."

A friend of yours who read the article says, “The study shows that sleep disorders lead to bullying in school children." Is this statement justified? If not, how best can you describe the conclusion that can be drawn from this study?

Association $\neq$ causation. Observational study, not experiment. Both sleep problems and bullying behaviors could be caused by other confounding variables

Section 1.5: Experiments

What are experiments?

- Researchers *assign* treatment to participants
- If treatment is *randomly* assigned, then it is a **randomized experiment**

Principles of experimental design

- **Control**: researchers control as much as possible. (This is the reason for placebos. Everyone takes a medication without knowing whether it contains active ingredients.)
- **Randomization**: “evens out” differences that are outside of the researchers’ control
- **Replication**: get a large sample!
- **Blocking**: similar to stratified sampling, where each block shares some characteristic, and treatments are randomly assigned within block
- **Blinding**: the people in the experiment don’t know which treatment group they are in. A study is **double-blind** if neither the people receiving nor the people giving the treatment know what the treatment is

Designing an Experiment

You would like to conduct an experiment to see if forensic scientists perform better at their jobs if they listen to music when they work. You want to determine if they work best without any music, with music that has no lyrics (instrumental), or with music that has lyrics. Briefly outline a design for this study.

Designing an Experiment

You would like to conduct an experiment to see if forensic scientists perform better at their jobs if they listen to music when they work. You want to determine if they work best without any music, with music that has no lyrics (instrumental), or with music that has lyrics. Briefly outline a design for this study.

- **Control**
- **Randomization**
- **Replication**
- **Blocking**
- **Blinding**

Designing an Experiment

You would like to conduct an experiment to see if forensic scientists perform better at their jobs if they listen to music when they work. You want to determine if they work best without any music, with music that has no lyrics (instrumental), or with music that has lyrics. Briefly outline a design for this study.

- **Control**: Control who listens to what music. Also, control group is no music
- **Randomization**: Randomly assign all forensic scientists in the lab who volunteer to one of the three groups
- **Replication**: make sure there are many people in each group
- **Blocking**: not required. could maybe block by age.
- **Blinding**: Not possible here. Everyone knows which music they are listening to.

Your Turn 1.5.1

A researcher is interested in the effects of exercise on mental health and he proposes the following study: Use stratified random sampling to ensure representative proportions of 18-30, 31-40 and 41- 55 year olds from the population. Next, randomly assign half the subjects from each age group to exercise twice a week, and instruct the rest not to exercise. Conduct a mental health exam at the beginning and at the end of the study, and compare the results.

1. What type of study is this?
2. What are the treatment and control groups in this study?
3. Does this study make use of blocking? If so, what is the blocking variable?
4. Does this study make use of blinding?
5. Comment on whether or not the results of the study can be used to establish a causal relationship between exercise and mental health, and indicate whether or not the conclusions can be generalized to the population at large.

Your Turn 1.5.1 (soln.)

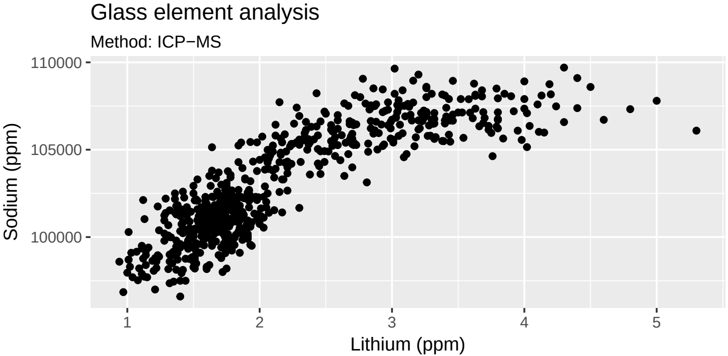
A researcher is interested in the effects of exercise on mental health and he proposes the following study: Use stratified random sampling to ensure representative proportions of 18-30, 31-40 and 41- 55 year olds from the population. Next, randomly assign half the subjects from each age group to exercise twice a week, and instruct the rest not to exercise. Conduct a mental health exam at the beginning and at the end of the study, and compare the results.

1. What type of study is this? Experiment
2. What are the treatment and control groups in this study? Treatment: excercise 2 times per week, Control: no exercise
3. Does this study make use of blocking? If so, what is the blocking variable? Yes, on age groups
4. Does this study make use of blinding? No. Participants know what group they are in.
5. Comment on whether or not the results of the study can be used to establish a causal relationship between exercise and mental health, and

Section 1.6: Looking at Numerical Data

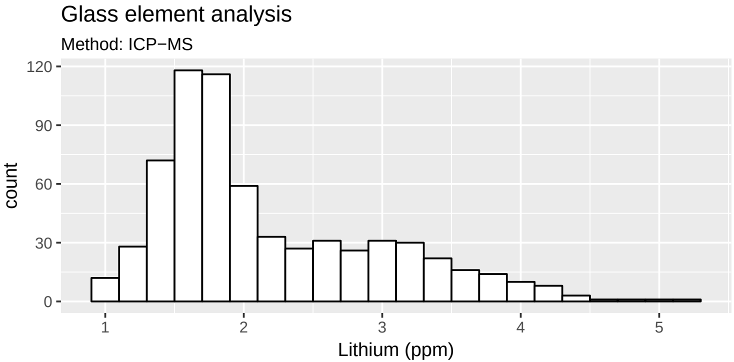
Scatterplots

- two numerical variables



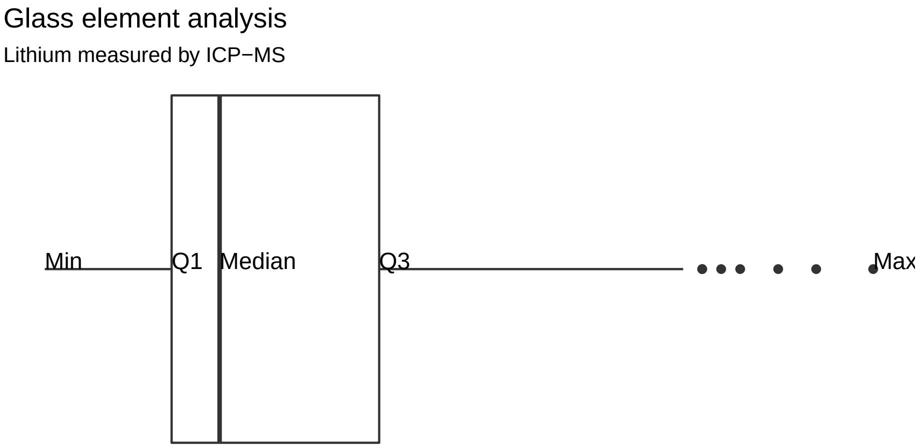
Histogram

- one numerical variable



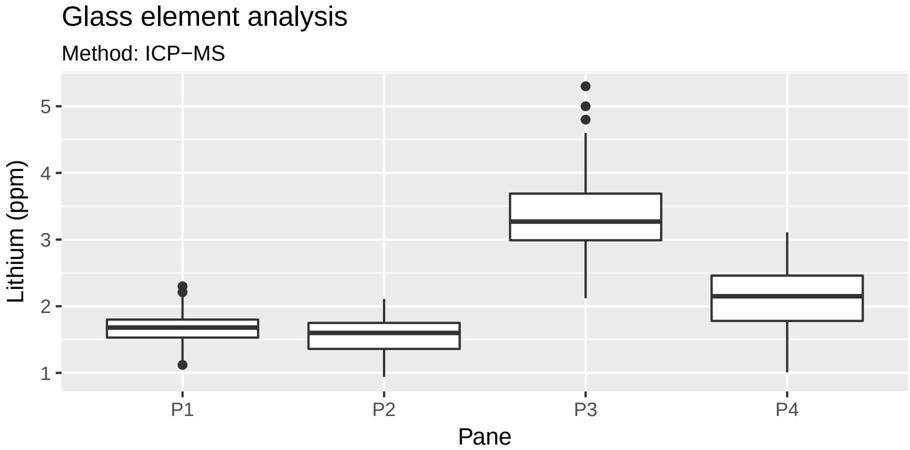
Boxplots

- assess distribution of one numerical variable



Boxplots

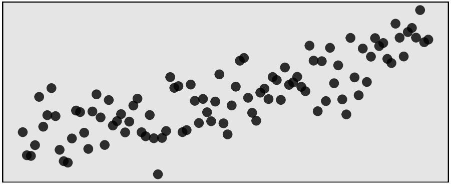
- compare distribution of a numerical variable across different values of a categorical variable



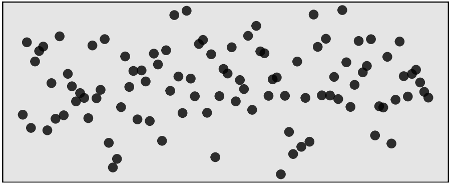
Your Turn 1.6.1

Indicate which plots show a positive association, a negative association, and no association. If there is an association, is it linear or non-linear?

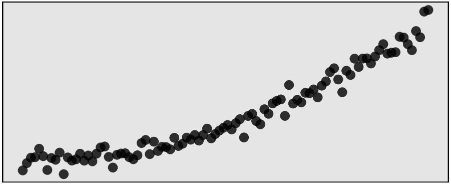
1



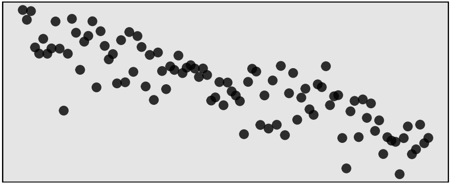
2



3



4



Your Turn 1.6.1 (soln.)

Indicate which plots show a positive association, a negative association, and no association. If there is an association, is it linear or non-linear?

- positive: 1 (linear), 3 (non-linear)
- negative: 4 (linear)
- no association: 2

Your Turn 1.6.2

For each part, compare distributions (1) and (2) based on their medians and IQRs.

1.
    1. 3, 5, 6, 7, 9
    2. 3, 5, 6, 7, 20
2.
    1. 3, 5, 6, 7, 9
    2. 3, 5, 7, 8, 9
3.
    1. 1, 2, 3, 4, 5
    2. 6, 7, 8, 9, 10
4.
    1. 0, 10, 50, 60, 100
    2. 0, 100, 500, 600, 1000

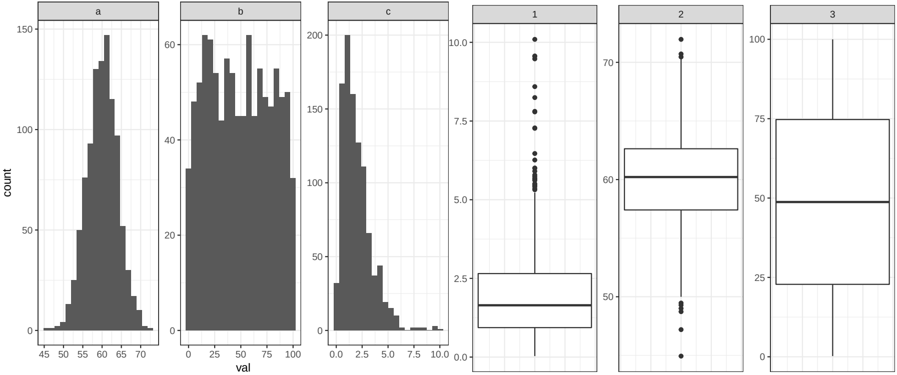
Your Turn 1.6.2 (soln.)

For each part, compare distributions (1) and (2) based on their medians and IQRs.

1. Very similar, outlier in (b) at 20
    1. 3, 5, 6, 7, 9 Med.=6, IQR $= 2$
    2. 3, 5, 6, 7, 20 Med.=6, IQR $= 2$
2. (b) tends to have higher values, has more variability
    1. 3, 5, 6, 7, 9 Med.=6, IQR $= 2$
    2. 3, 5, 7, 8, 9 Med.=7, IQR $= 3$
3. Same spread, (b) higher than (a)
    1. 1, 2, 3, 4, 5 Med.=3, IQR $= 2$
    2. 6, 7, 8, 9, 10 Med.=8, IQR $= 2$
4. Both have median equal to IQR. (a) much smaller than (b)
    1. 0, 10, 50, 60, 100 Med.=50, IQR $= 50$
    2. 0, 100, 500, 600, 1000 Med.=500, IQR $= 500$

Your Turn 1.6.3

Describe the data in the histograms and match them to the boxplots.



Your Turn 1.6.3 (soln.)

Histogram	Boxplot
a	2
b	3
c	1

Section 1.7: Looking at Categorical data

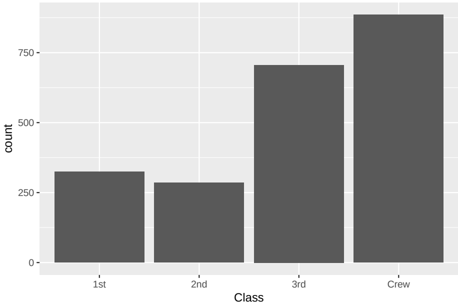
Contingency tables

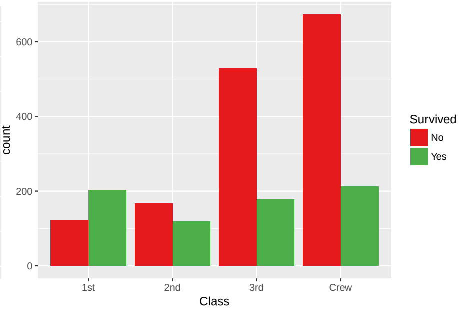
- contigency tables show observed counts of categorical variables
- Example: `titanic` data

Class	Died	Survived	Total
1st	122	203	325
2nd	167	118	285
3rd	528	178	706
Crew	673	212	885
Total	1490	711	2201

Bar charts

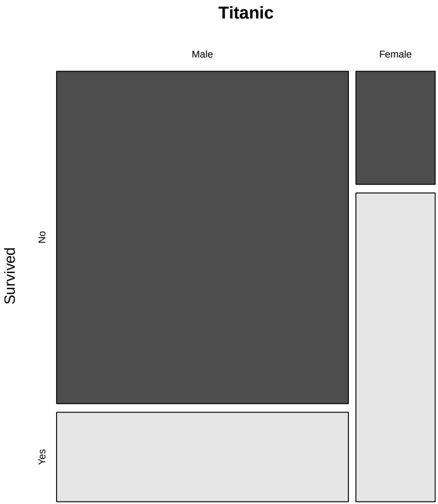
- show counts of different groups in the data





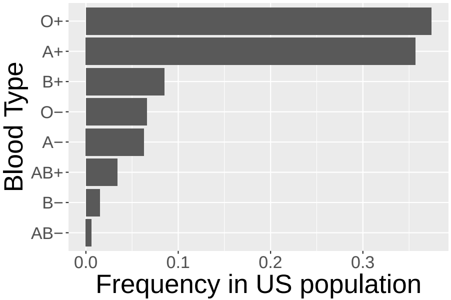
Mosaic Plots

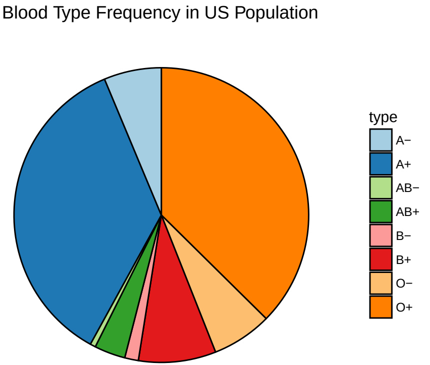
- plots to compare two categorical variables



Your Turn 1.7.1

The bar plot and the pie chart below show the ditribution of blood types in the US.




Blood Type Frequency in US Population

Your Turn 1.7.1 (cont.)

1. What features are apparent in the bar plot but not in the pie chart?
2. What features are apparent in the pie chart but not in the bar plot?
3. Which graph would you prefer to use for displaying these categorical data?

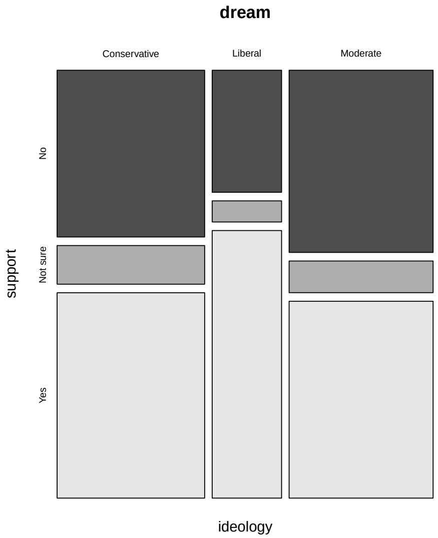
Your Turn 1.7.1 (soln.)

1. What features are apparent in the bar plot but not in the pie chart? We see the order of the categories and the relative frequencies in the bar plot.
2. What features are apparent in the pie chart but not in the bar plot? There are no features that are apparent in the pie chart but not in the bar plot.
3. Which graph would you prefer to use for displaying these categorical data? We usually prefer to use a bar plot as we can also see the relative frequencies of the categories in this graph.

Your Turn 1.7.2

A random sample of US registered voters were asked if they support the DREAM Act. The survey also asked about the political ideology of the respondents. Based on the mosaic plot shown on the next slide, do views on the DREAM Act and political ideology appear to be **independent** ? Explain.

Your Turn 1.7.2 (cont.)



Your Turn 1.7.2 (soln.)

A random sample of US registered voters were asked if they support the DREAM Act. The survey also asked about the political ideology of the respondents. Based on the mosaic plot shown on the next slide, do views on the DREAM Act and political ideology appear to be **independent** ? Explain.

The vertical locations at which the ideological groups break into the Yes, No, and Not Sure categories differ, which idicates that likelihood of supporting the DREAM act varies by political ideology. This suggests that the two variables may be dependent.